Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents

Oana Cocarascu
Department of Computing,
Imperial College London
London, UK
oc511@imperial.ac.uk

Antonio Rago
Department of Computing,
Imperial College London
London, UK
a.rago15@imperial.ac.uk

Francesca Toni
Department of Computing,
Imperial College London
London, UK
ft@imperial.ac.uk

ABSTRACT

The aggregation of online reviews is fast becoming the chosen method of quality control for users in various domains, from retail to entertainment. Consequently, fair, thorough and explainable aggregation of reviews is increasingly sought-after. We consider the movie review domain, and in particular Rotten Tomatoes' ubiquitous (and arguably over-simplified) aggregation method, the Tomatometer Score (TS). For a movie, this amounts to the percentage of critics giving the movie a positive review. We define a novel form of argumentative dialogical agent (ADA) for explaining the reasoning within the reviews. ADA integrates: 1.) NLP with reviews to extract a Quantitative Bipolar Argumentation Framework (QBAF) for any chosen movie to provide the underlying structure of explanations, and 2.) gradual semantics for QBAFs for deriving a dialectical strength measure for movies, as an alternative to the TS, satisfying desirable properties for obtaining explanations. We evaluate ADA using some prominent NLP methods and gradual semantics for QBAFs. We show that they provide a dialectical strength which is comparable with the TS, while at the same time being able to provide dialogical explanations of why a movie obtained its strength via interactions between the user and ADA.

KEYWORDS

Dialogical Interactions; Explainability; Argument Mining; Quantitative Argumentation

ACM Reference Format:
Oana Cocarascu, Antonio Rago, and Francesca Toni. 2019. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

In an age in which e-commerce and audio/video streaming are dominant markets for consumers, products' online reviews are fast becoming the preferred method of quality control for users. The aggregation of these reviews allows users to check the quality of a product while avoiding reviews which may be incoherent, irrelevant or even malicious. Users are increasingly turning to aggregation sites that give an overview of reviews, e.g. *Metacritic*'s[1] aggregation of trusted critic reviews for music albums. The method of aggregation and the way in which it is presented to a user (i.e. an *explanation* of the aggregation) must be thorough, trustworthy and intuitive in order to both satisfy existing users and attract new ones. Within the movie domain, *Rotten Tomatoes*[2] (RT) is a popular review site that aggregates critics' reviews (simplified to a binary classification of either *Fresh* or *Rotten*) to obtain an overall percentage of critics who like the movie and critics who do not, the *Tomatometer Score* (TS). The TS is further simplified to a binary classification for the movie of Fresh or Rotten once again, based on whether it is greater or equal to 60% or not, respectively. A short consensus is also written by a moderator to give a linguistic summary of the reviews. This simplification into TS, fresh/rotten classification and consensus gives users a quick way to determine whether a movie is worth watching or not. The simple and recognisable TS has been subsequently incorporated into streaming sites (e.g. *iTunes*), search engines (e.g. *Google*) and ticket sales apps (e.g. *Fandango*).

This phenomenon is not without its problems. Within the movie industry numerous contributors have bemoaned RT's apparently detrimental effect on the industry[3], with one of the more plausible claims being that the ubiquitous TS oversimplifies a movie's aggregated review and "hacks off critical nuance"[4]. Another issue is that the TS score represents the percentage of top critics who felt anywhere from mildly to wildly positively about a given movie. This means that a critic's mixed review that is slightly positive overall will have the same weight as a rave review from another critic, leading to the case where a movie with a maximum TS could be composed of only generally positive reviews. Also, the TS does not take into account user preferences and so factors which decrease the TS may not have any relevance in a user's personal selection criteria, meaning movies may be overlooked when they may actually be perfectly suited to a user's tastes. Taking into account user preferences would raise a plethora of privacy concerns unless a method to *explain* the aggregation sufficiently for all users is undertaken, so that they can decide for themselves. That being said, the percentage of critics who believe a film is *fresh* rather than *rotten* is a useful and intuitive indicator for users of all backgrounds. Therefore, if the TS or a similar measure were supplemented with dialogical explanations empowering users to interact with the system for more information about a movie's aggregated review, this may alleviate the issues stated above while maintaining the advantages of the TS.

We propose a novel model for an argumentative dialogical agent (ADA), overviewed in Figure 1, that can extract explanations

[1] https://www.metacritic.com

[2] https://www.rottentomatoes.com
[3] https://www.theatlantic.com/amp/article/543090/
[4] https://nyti.ms/2xcXS0Y

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

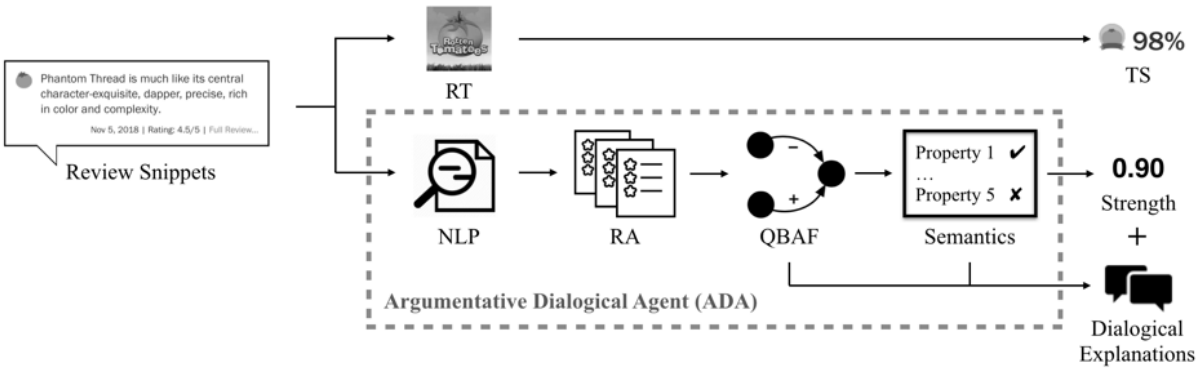

Figure 1: An overview of ADA for extracting aggregated scores with dialogical explanations from (snippets from) reviews, compared with the original RT method of computing a TS score as an aggregation method.

from review snippets and can engage human users within dialogical exchanges about the TS for movies, using these explanations. ADA relies upon a *feature*-based conceptualisation of reviews (for movies, taken as example products in this paper) and mines *review aggregations* (RAs) from snippets drawn from reviews by *critics* to obtain *votes* (on both movies and their features) that are in turn used to generate Quantitative Bipolar Argumentation Frameworks (QBAFs). These are argumentation frameworks where arguments can attack or support arguments, and where arguments are equipped with a *base score*, as in [4], and a *strength* obtained using a gradual semantics. We select QBAFs as a principle formalism for our method firstly for their suitability in this application. For a given movie or one of its features: the base score gives a measure of the reviews' general sentiment towards it, while the strength measure takes into account the strengths of related features, paving the way to dialogical explanations based on these dialectical relations. QBAFs also generalise a number of other argumentation frameworks (see [4]) and so could be restricted as such in this application if required. We mine RAs from snippets using Natural Language Processing (NLP) methods, namely Argument Mining and Sentiment Analysis. ADA uses NLP to determine whether snippets provide positive or negative votes for (features of) movies, by looking for arguments supporting or attacking, respectively, the (features of) movies. ADA then uses the obtained QBAFs to extract dialogical explanations for the strengths of the movies (which correlate with their TS) obtained via gradual semantics of these frameworks that exhibit properties conducive to the computed strength to mimic the TS.

2 RELATED WORK

Argument Mining (AM) is a recent but well studied field (see [21] for a recent overview). It can be seen as an advancement of Sentiment Analysis (SA) and Opinion Mining (as indicated in [18, 21]). Whilst the goal of SA and Opinion Mining is to identify what users think, the goal of AM is to understand the reasons why users think the way they do, not limiting to opinions and their sentiment polarity. In our setting, one can see opinions as arguments as to why users should or should not choose a product (e.g. watch a movie).

Amongst several approaches to AM, the one we use is closest to that in [8], where *attack* and *support* relations are mined from tweets, i.e. texts which are similar in size to the review snippets we consider in this paper. In our experimental evaluation, we use two different techniques for supporting AM: one concerning feature-based SA and the other based on deep learning [11].

Several works with applications in the movie setting make use of (feature-based) SA and machine learning techniques. Related works include those proposing aggregation methods for recommending movies [36] and document-level SA in the movie setting [6, 10, 23, 34, 35]. Other works focus on extracting important features from reviews using machine learning techniques [20, 25, 37].

Our dialogical explanations can be seen as an argumentation-based summary of reviews. [15] propose a general summarisation framework based on abstract argumentation used to select sentences from text. This summarisation extracts the most relevant information (arguments) from reviews. We use the mined arguments and relations to generate votes that are used to obtain argumentation frameworks. Argumentation has also been used extensively to generate explanations in AI: e.g. [16] formalise dialectical explanations for argument-based reasoning, while in [30], an argumentation-driven recommender system provides explanations for recommendations extracted from argumentation frameworks. Our argumentative agents extract the argumentation frameworks from votes in turn obtained by NLP, in an innovative fashion.

Dialogical argumentation has been studied in various settings, e.g. dialogue games for argumentation [28], persuasion dialogue [29] and strategic argumentation [19]. We use it for supporting the exchange of explanations with users. Argumentation has been used to support dialogical agents in games [3], negotiation [22] and decision making with goals [14]. Separately from argumentation, dialogue is an important medium for agents with other purposes e.g. for task completion [33] or information access [12].

3 BACKGROUND

As in [4], a QBAF is a quadruple $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ consisting of a set $\mathcal{A}$ of arguments, a binary (*attack*) relation $\mathcal{L}^-$ on $\mathcal{A}$, a binary (*support*) relation $\mathcal{L}^+$ on $\mathcal{A}$ and a total function $\tau : \mathcal{A} \to \mathbb{I}$; for any $\alpha \in \mathcal{A}$, $\tau(\alpha)$ is the *base score* of α. Throughout this paper, we will use $\mathbb{I} = [0, 1]$. For any $\alpha \in \mathcal{A}$, the *strength* of α is given by $\sigma(\alpha)$, where $\sigma : \mathcal{A} \to \mathbb{I}$ is a total function (also referred to as a *gradual semantics*). For any

set of arguments $S \subseteq \mathcal{A}$ we denote $\sigma(S)$ a sequence (in any order) of all elements of the set $\{\sigma(\alpha)|\alpha \in S\}$ (thus $\sigma(S) \in \mathbb{I}^*$).

In our experimental evaluation, we consider three gradual semantics for a QBAF: *QuAD* [5], *DF-QUAD* [32] and the *Restricted Euler-based semantics* (REB) [1]. We briefly recap them, using the following notation: for a given QBAF $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ and $\alpha \in \mathcal{A}$: $\mathcal{L}^-(\alpha) = \{\beta \in \mathcal{A}|(\beta, \alpha) \in \mathcal{L}^-\}$ is the set of *attackers* of α; $\mathcal{L}^+(\alpha) = \{\beta \in \mathcal{A}|(\beta, \alpha) \in \mathcal{L}^+\}$ is the set of *supporters* of α.

In QuAD, for $\alpha \in \mathcal{A}$, $\sigma(\alpha) = g(\tau(\alpha), \mathcal{F}_a(\tau(\alpha), \sigma(\mathcal{L}^-(\alpha))), \mathcal{F}_s(\tau(\alpha), \sigma(\mathcal{L}^+(\alpha))))$ where if $(\alpha_1, \ldots, \alpha_n)$ is an arbitrary permutation of the $(n \geq 0)$ attackers in $\mathcal{L}^-(\alpha)$, $\sigma(\mathcal{L}^-(\alpha)) = (\sigma(\alpha_1), \ldots, \sigma(\alpha_n))$ (similarly for supporters). The operator $g: \mathbb{I} \times \mathbb{I} \cup \{nil\} \times \mathbb{I} \cup \{nil\} \to \mathbb{I}$ is defined, for $v_0, v_a, v_s \in \mathbb{I}$, as: if $v_s = nil$ and $v_a \neq nil$, $g(v_0, v_a, v_s) = v_a$; if $v_a = nil$ and $v_s \neq nil$, $g(v_0, v_a, v_s) = v_s$; if $v_a = v_s = nil$, $g(v_0, v_a, v_s) = v_0$; otherwise $g(v_0, v_a, v_s) = \frac{v_a + v_s}{2}$. Letting $\times$ stand for either a or s, the operator $\mathcal{F}_\times$ is defined as $\mathcal{F}_\times: \mathbb{I}^* \to \mathbb{I}$, where for $S = (v_1, \ldots, v_m) \in \mathbb{I}^*, (w_1, \ldots, w_n)$ is an arbitrary permutation of the non-zero elements in S[5]: if $n = 0$: $\mathcal{F}_\times(v_0, S) = nil$; if $n = 1$: $\mathcal{F}_\times(v_0, S) = f_\times(v_0, w)$; if $n > 1$: $\mathcal{F}_\times(v_0, (w_1, \ldots, w_n)) = f_\times(\mathcal{F}_\times(v_0, (w_1, \ldots, w_{n-1})), w_n)$; with the *base expressions* $f_\times: \mathbb{I} \times \mathbb{I} \to \mathbb{I}$ defined, for $v_0, v \in \mathbb{I}$, as: $f_a(v_0, v) = v_0 \cdot (1 - v)$ and $f_s(v_0, v) = v_0 + v - v_0 \cdot v$.

In DF-QuAD, for any $\alpha \in \mathcal{A}$ with $\tau(\alpha) = v_0$ and n attackers with strengths $v_1, \ldots, v_n$ and m supporters with strengths $v'_1, \ldots, v'_m$, $\sigma(\alpha) = \mathcal{C}(v_0, \mathcal{F}(v_1, ..., v_n), \mathcal{F}(v'_1, ..., v'_m))$. The combination function $\mathcal{C}$, for an argument with base score v_0, attackers with strengths $v_1, ..., v_n$ (for $n \geq 0$, $n = 0$ amounts to the argument having no attackers) and supporters with strengths $v'_1, ..., v'_m$ (for $m \geq 0$, $m = 0$ amounts to the argument having no supporters) is defined as follows, for $v_a = \mathcal{F}(v_1, ..., v_n)$ and $v_s = \mathcal{F}(v'_1, ..., v'_m)$: if $v_a = v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0$; else if $v_a > v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0 - (v_0 \cdot |v_s - v_a|)$; otherwise $\mathcal{C}(v_0, v_a, v_s) = v_0 + ((1 - v_0) \cdot |v_s - v_a|)$. Given n arguments with strengths $v_1, ..., v_n$, if $n = 0$ then $\mathcal{F}(v_1, ..., v_n) = 0$, otherwise $\mathcal{F}(v_1, ..., v_n) = 1 - \prod_{i=1}^{n}(|1 - v_i|)$.

The REB semantics is such that for $\alpha \in \mathcal{A}$, $\sigma(\alpha) = 1 - \frac{1 - \tau(\alpha)^2}{1 + \tau(\alpha) \cdot e^E}$ where $E = \sum_{\beta \in \mathcal{L}^+(\alpha)} \sigma(\beta) - \sum_{\gamma \in \mathcal{L}^-(\alpha)} \sigma(\gamma)$.

In general, the choice of a semantics for an application is based on the desirable behaviour it should exhibit, which can be defined in the form of properties the semantics should satisfy, such as the properties of *(strict) balance* ((S)B) and *(strict) monotonicity* ((S)M) [4]. The former two state that an imbalance between the strengths of an argument's attackers and supporters must correspond to a difference in its strength and its base score, whereas the latter two require that the strength of an argument depends monotonically on its base score and the strengths of its attackers and supporters and their strengthening/weakening will do likewise to the argument. The three semantics we consider hold different combinations of the four mentioned properties, as shown in Table 1 [4]. Note that the SB property is not satisfied by any of the gradual semantics considered here. This is not a concern for the setting considered in this paper as SB states that any difference in the base score and strength of an argument must correspond to a specific form of *dominance* between the sets of attackers and supporters, a requirement which is not desirable here, e.g. given an argument with two attackers with strengths 0.1 and 0.2 and a single supporter with strength 0.9, we may require that it has a strength higher than its base score, which violates SB (see [4] for details).

Semantics	Properties			
	B	SB	M	SM
QuAD	·	·	✓	·
DF-QuAD	✓	·	✓	·
REB	✓	·	✓	✓

Table 1: QBAF semantics and properties of *(strict) balance* ((S)B) and *(strict) monotonicity* ((S)M) from the literature.

4 ARGUMENTATIVE DIALOGICAL AGENT

ADA is designed around a *feature-based characterisation* of movies[6]:

Definition 4.1. Let $\mathcal{M}$ be a given set of movies, and $m \in \mathcal{M}$ be any movie. A *feature-based characterisation of* m is a finite set $\mathcal{F}$ of *features* with *sub-features* $\mathcal{F}' \subset \mathcal{F}$ such that each $f' \in \mathcal{F}'$ has a unique *parent* $p(f') \in \mathcal{F}$; for any $f \in \mathcal{F} \backslash \mathcal{F}'$, we define $p(f) = m$.

A sub-feature is more specific than its parent feature, e.g. for the movie m = *Wonder Wheel*, a feature may be *acting*, the parent of the sub-feature *Kate Winslet*. Below we will often refer to elements of $\mathcal{F} \backslash \mathcal{F}'$ as features, and to elements of $\mathcal{F}'$ as sub-features. Also, we will refer to a sub-feature with parent f as a sub-feature of f.

This feature-based characterisation may be obtained from metadata or the top critics' snippets that appear on RT movie pages (e.g. see *Wonder Wheel*'s reviews[7]). In doing so, for *Wonder Wheel*, we may obtain features $\{f_A, f_D, f_W, f_T\}$, where f_A is *acting*, f_D is *directing*, f_W is *writing* and f_T is *themes*.[8] The sub-features in $\mathcal{F}'$ may be of different types, namely *single* (e.g. for features f_D or f_W, if we only consider movies with a single director or writer) or *multiple* (e.g. for f_A, as movies will generally have more than one actor (*Wonder Wheel* has *Jim Belushi*, *Justin Timberlake* and *Kate Winslet* as sub-features of f_A), and f_T, since movies will generally be associated with several themes). Single sub-features can be equated with the feature (e.g. for *Wonder Wheel*, *Woody Allen* is the sole director and so this sub-feature can be represented by f_D itself). Furthermore, sub-features may be *predetermined*, namely obtained from meta-data (as for the sub-features with parents f_A, f_D, f_W in the running example), or *mined* from (snippets of) reviews (e.g. for *Wonder Wheel* the sub-feature *amusement park* of f_T may be mined rather than predetermined). To determine the mined sub-features of a movie we can use semantic information, e.g. in our experiments in Section 5.1 we use the semantic network ConceptNet[9] to identify related terms for f_T. For example, for *Wonder Wheel*, we identify the sub-feature *amusement park* (f'_{T1}) as several reviews mention the related terms *Coney Island* and *fairground*, as in *'like the fairground ride for which it's named, Wonder Wheel is entertaining'*.

Using this feature-based characterisation of a movie and snippets from the movie reviews by critics, as in RT, ADA uses

[5] This formulation is a modification of the original formulation of $\mathcal{F}_\times$, in which $(w_1, \ldots, w_n)$ was not used.

[6] Movies are our chosen products here, but ADA can be defined for any products.

[7] https://www.rottentomatoes.com/m/wonder_wheel/reviews/?type=top_critics

[8] In the experiments in Section 5.1 we limited the analysis to exactly these four features only as these are the ones that occur most frequently in the movie domain.

[9] http://conceptnet.io/

NLP to generate votes on arguments, amounting to the movie in question and its (sub-)features. The result is a *review aggregation* for the movie, defined in Section 4.1, which the agent then transforms into a QBAF, as defined in Section 4.2, and from which the agent generates *dialogical explanations*, as defined in Section 6.

4.1 Extracting Review Aggregations

Let $m \in \mathcal{M}$ be any movie and $\mathcal{F}$ be a feature-based characterisation of m as given in Definition 4.1. Let $\mathcal{A}$ denote $\{m\} \cup \mathcal{F}$, referred to as the set of *arguments*. We then define the following:

Definition 4.2. A *review aggregation* for m is a triple $\mathcal{R}(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ where:

- $\mathcal{C}$ is a finite, non-empty set of *critics*;
- $\mathcal{V} : \mathcal{C} \times \mathcal{A} \rightarrow \{-, +\}$ is a partial function, with $\mathcal{V}(c, \alpha)$ representing the *vote* of critic $c \in \mathcal{C}$ on argument $\alpha \in \mathcal{A}$.

Straightforwardly, a positive/negative vote from a critic on a (sub-)feature of the movie signifies positive/negative stance on that (sub-)feature and a positive/negative vote on m signifies positive/negative stance on the overall movie.

A review aggregation can be *augmented* by exploiting the parent relation. Indeed, a vote for/against an argument can be seen as a vote for/against the argument's parent.

Definition 4.3. Given a review aggregation $\mathcal{R}_0(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V}_0 \rangle$, an *augmented review aggregation* $\mathcal{R}(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ is such that for any $c \in \mathcal{C}$ and any $\alpha \in \mathcal{A}$:

- if $\mathcal{V}_0(c, \alpha)$ is defined, then $\mathcal{V}(c, \alpha) = \mathcal{V}_0(c, \alpha)$; else
- if $|\{\beta \in \mathcal{A} | p(\beta) = \alpha \wedge \mathcal{V}_0(c, \beta) = +\}| > |\{\gamma \in \mathcal{A} | p(\gamma) = \alpha \wedge \mathcal{V}_0(c, \gamma) = -\}|$ then $\mathcal{V}(c, \alpha) = +$; else
- if $|\{\beta \in \mathcal{A} | p(\beta) = \alpha \wedge \mathcal{V}_0(c, \beta) = +\}| < |\{\gamma \in \mathcal{A} | p(\gamma) = \alpha \wedge \mathcal{V}_0(c, \gamma) = -\}|$ then $\mathcal{V}(c, \alpha) = -$; else
- $\mathcal{V}(c, \alpha)$ is undefined.

For example, let c's vote on f_A be undefined, c's vote on sub-feature f'_{A1} of f_A be + and there be no − votes from c on any other sub-features of f_A. We then assume that c's overall stance on f_A is positive and therefore set c's vote on f_A to +. This notion of augmenting the review aggregation combats the brevity of the snippets causing the review aggregation being too sparsely populated.

ADA uses NLP to extract review aggregations, from which augmented review aggregations are obtained as in Definition 4.3. The NLP is used to analyse each critic's review independently, tokenising each into sentences, which are then split into phrases when specific keywords (*but, although, though, otherwise, however, unless, whereas, despite*) occur. Each phrase may then constitute an argument with a vote from its critic in the review aggregation. For illustration, consider the following review for m = *Wonder Wheel* from critic c_1: *Allen, 82, has his ups and downs, and while there have been more downs than ups lately he is always worth the benefit of the doubt. But Wonder Wheel is a ride to nowhere.* ADA extracts two phrases: p_1: *Allen...doubt.* and p_2: *Wonder...nowhere.* A review comprising a single sentence, e.g. c_2: *Despite a stunning performance by Winslet and some beautiful cinematography by Vittorio Storaro, Wonder Wheel loses its charms quickly and you'll soon be begging to get off this particular ride* may be split into p_3: *Despite...Storaro.* and p_4: *Wonder...ride.* Finally, consider the review from critic c_3: *As we watch Allen worry and nitpick over the way women fret over aging, painting ginny as pathetic, jealous, insecure, and clownish, it's dull, unoriginal, and offensive. Frankly, we've had enough Woody Allen takes on this subject.* Here, ADA extracts two phrases concerning *Woody Allen*: p_5: *As ... offensive* and p_6: *Frankly ... subject.*

In order to determine the arguments on which the votes acts, ADA uses a *glossary* G using movie-related words for each feature as well as for movies in general. G is as follows (for any $m \in \mathcal{M}$): $G(m) = \{movie, film, work\}$; $G(f_D) = \{director\}$; $G(f_A) = \{acting, cast, portrayal, performance\}$; $G(f_W) = \{writer, writing, screenplay, screenwriter, screenwriting, storyline, script, character\}$. When determining the argument on which a vote acts, sub-features take precedence over features. A mention of "Kate Winslet" (f'_{A1}) (with or without a word from $G(f_A)$) connects with f'_{A1}, whereas a sole mention of any word from $G(f_A)$ connects with f_A. A text that contains two entities (a sub-feature or a word from the glossary) corresponding to different (sub-)features results in two arguments (and votes), one for each (sub-)feature identified.

Once arguments have been identified, votes on them can also be extracted by NLP techniques. In our experiments in Section 5.1, we experiment with and compare two alternative NLP techniques to determine votes and thus conclude the mining of review aggregations. These techniques are Sentiment Analysis (SA) and identifying attack/support by Argument Mining (AM), as described next.

4.1.1 Mining votes using SA. The sentiment polarity of each phrase is translated into a (negative or positive) vote from the corresponding critic. We impose a threshold on the sentiment polarity to filter out phrases that can be deemed to be "neutral" and therefore cannot be considered to be votes. Votes are then assigned to arguments based on occurrences of words from G. For example, ADA may deem the phrase p_1 to have a polarity of 0.611 and to be assigned to f_D, therefore giving $\mathcal{V}(c_1, f_D) = +$. Similarly, p_2 may result in a sentiment/target argument pair of $(-0.518, m)$, therefore giving $\mathcal{V}(c_1, m) = -$. If the neutrality threshold is ± 0.6, a positive vote corresponding to p_1 is assigned to f_D, whereas the negative vote corresponding to p_2 is not assigned to m. From the review from c_2, ADA may extract one vote for the sub-feature *Kate Winslet* (f'_{A1}) and one for the movie in general, i.e. p_3 gives $(0.833, f'_{A1})$ therefore $\mathcal{V}(c_2, f'_{A1}) = +$, while p_4 gives $(-0.604, m)$ therefore $\mathcal{V}(c_2, m) = -$. It should be noted that if the feature *cinematography* had been included in our $\mathcal{F}$ then we may have had another vote from c_2. This could be achieved by using more metadata of the movies and hence an occurrence of *Storaro* would correspond to a vote on cinematography. We could also identify more features by using topic modelling and aggregate similar terms to identify the clusters that lead to the top features (similar to the four we selected). However, in the RT setting, cinematography is not a topic which is commented on most often in review snippets. If the review of a single critic results in several phrases associated with an argument with different polarities, ADA takes that with the highest sentiment magnitude to determine the vote on that argument. For example, given: p_5: $(-0.659, f_D)$ and p_6: $(0.500, f_D)$, p_5 supersedes and $\mathcal{V}(c_3, f_D) = -$. Here, the sentiment for p_6 is incorrect but the neutrality threshold leads to it being ignored. ADA determines the votes for the mined f_T in the same way as the other features. For example, given

p_7: *like the fairground ride for which it's named, Wonder Wheel is entertaining* leading to $(0.741, f'_{T1})$, ADA obtains $\mathcal{V}(c, f'_{T1}) = +$.

4.1.2 Mining votes using AM. Relation-based AM, as in [11], can be used for identifying attack/support/neither relations between two extracted phrases, where attack leads to a negative vote and support to a positive vote. For example, consider the feature f_A and a sentence mentioning acting. AM may determine whether there is an argumentative relation between the sentence and the pre-set argument f_A, which may be read as the natural language argument: *the acting was good*, and this relation can be used to assign a vote. For illustration, considering the phrases extracted from critic c_1, AM may give that p_1 supports f_D and p_2 attacks m, and therefore $\mathcal{V}(c_1, f_D) = +$ and $\mathcal{V}(c_1, m) = -$. Neutrality (of the form discussed for SA) is given by the third relation (neither attack not support).

4.2 Formulating QBAFs from Review Aggregations

In order to obtain a QBAF from a review aggregation, we must determine: the arguments, their base scores and between which arguments attacks and supports are present. For arguments we choose $\mathcal{A}$ as before, for base scores we use an intuitive aggregation of critics' votes, and for the attacks and supports, straightforwardly, we impose that a (sub-)feature attacks or supports its parent argument depending on its aggregated stance, as follows:

Definition 4.4. Let $\mathcal{R}(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ be any (augmented) review aggregation for $m \in \mathcal{M}$. For any $\gamma \in \mathcal{A} = \mathcal{F} \cup \{m\}$, let $\mathcal{V}^+(\gamma) = |\{c \in \mathcal{C} | \mathcal{V}(c, \gamma) = +\}|$ and $\mathcal{V}^-(\gamma) = |\{c \in \mathcal{C} | \mathcal{V}(c, \gamma) = -\}|$. Then, the *QBAF corresponding to* $\mathcal{R}(m)$ is $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ such that

$$\mathcal{L}^- = \{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) > \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)\} \cup$$
$$\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) < \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) > \mathcal{V}^-(\alpha)\} \cup$$
$$\{(\alpha,m) | \alpha \in \mathcal{F} \wedge m = p(\alpha) \wedge \mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)\};$$
$$\mathcal{L}^+ = \{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) \geq \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)\} \cup$$
$$\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) \leq \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) \leq \mathcal{V}^-(\alpha)\} \cup$$
$$\{(\alpha,m) | \alpha \in \mathcal{F} \wedge m = p(\alpha) \wedge \mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)\};$$

$$\tau(m) = 0.5 + 0.5 \cdot \frac{\mathcal{V}^+(m) - \mathcal{V}^-(m)}{|\mathcal{C}|} \text{ and } \forall f \in \mathcal{F},\ \tau(f) = \frac{|\mathcal{V}^+(f) - \mathcal{V}^-(f)|}{|\mathcal{C}|}.$$

An attack is defined as either from a feature with dominant negative votes (with respect to positive votes) towards the movie itself or from a sub-feature with dominant negative (positive) votes towards a feature with dominant positive (negative, respectively) votes. The latter type of attack can be exemplified by a sub-feature of f_A with positive stance attacking the negative (due to other votes/arguments) feature f_A, which attacks m. Conversely, a support is defined as either from a feature with dominant positive votes towards the movie itself or from a sub-feature with dominant positive (negative) votes towards a feature with dominant positive (negative, respectively) votes. The latter type of support can be exemplified by a sub-feature of f_A with negative stance supporting the negative feature f_A, which attacks m. It should be noted that (sub-)features with equal positive and negative votes are treated as supporters, though we could have assigned no relation here.

The base score of m, $\tau(m) \in [0, 1]$, has been adapted from [31], where several useful properties thereof are shown. Intuitively, $\tau(m) = 1$ represents all critics having a positive stance on the movie while $\tau(m) = 0$ requires universally negative stance. The base score of a (sub-)feature f is again in $[0, 1]$ where, differently to movies since a feature already represents positive/negative sentiment towards the argument it supports/attacks, $\tau(f) = 0$ represents no dominant negative/positive stance from the critics on f while $\tau(f) = 1$ represents universally negative/positive stance on f.

5 RESULTS & DISCUSSION

In this section, we evaluate empirically (Section 5.1) and qualitatively (Section 5.2) different realisations of ADA. In particular, we consider which method for vote mining (amongst SA and AM) and which gradual semantics for QBAFs (amongst those detailed in Section 3) allow for better correlation with the TS. Moreover, we identify properties of gradual semantics that render them better suited for generating explanations from QBAFs.

5.1 Experimental Results

We tested our proposed method on box-office movies from January 2015 to August 2018 inclusive and on the top 100 movies of all time[10], giving a total of 1281 movies after removing those without reviews. We conducted experiments to determine the performance of our predicted ratings compared to the TS using both SA and AM techniques. We report the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) on the scale 0-100, along with scaling the score to a 0-9 point scale to reflect the 'grade' of the movie (e.g. 7 = [70, 79]). We also show the classification report of predicting whether a movie is certified Fresh or Rotten, where movies with TS ≥ 60 are classified as Fresh and those with TS < 60 as Rotten. In all of our experiments, we analyse movies for which we obtain votes from at least 33% of the critics who reviewed the movies, removing those for which the NLP techniques mined few votes, resulting in approximately 900 movies and 35219 snippets (which gave 840874 tokens). From these, we consider 33554 arguments that are analysed using the method outlined in Section 4.

The first experiment uses SA, for which we deployed an off-the-shelf classifier on movie reviews [24]. The second experiment uses AM, for which we deployed the dataset and the basis of the best performing neural model from [11]. This Deep Learning architecture consists of two parallel Long-Short-Term-Memory networks (LSTMs) that model the two texts (phrases) separately. We concatenated the two outputs of the LSTMs and fed them to a *softmax* classifier to determine the argumentative relation (*support*, *attack* or *neither*). We used a 100-dimensional GloVe embedding [27] and limited the input texts to 30 words. We set the LSTM dimension to 32 and the activation function to *tanh*. We applied dropout of 0.2 after the LSTMs. These hyper-parameters are those which led to the best performing model. The other options for hyper-parameters (e.g. dropout and LSTM 64, 100, 128) led to a less performant model for detecting argumentative relations, in both accuracy and F1 measure. We trained for 1000 epochs with patience set to 5 with 20% of the dataset as the validation set.

[10] https://www.rottentomatoes.com/top/bestofrt/

		% Scale			Grade Scale		
		MAE	RMSE	PCC	MAE	RMSE	PCC
SA	QuAD	16.004	19.306	0.731	1.535	1.925	0.708
	DF-QuAD	15.978	19.329	0.729	1.527	1.927	0.707
	REB	17.530	20.296	0.752	1.693	2.005	0.725
AM	QuAD	16.404	19.662	0.714	1.582	1.960	0.689
	DF-QuAD	16.343	19.677	0.712	1.573	1.961	0.687
	REB	18.125	20.911	0.734	1.735	2.045	0.712

Table 2: Errors between strength and TS for SA and AM.

			Precision	Recall	F_1
SA	QuAD	Fresh	0.84	0.93	0.88
		Rotten	0.80	0.62	0.70
	DF-QuAD	Fresh	0.84	0.94	0.88
		Rotten	0.81	0.59	0.68
	REB	Fresh	0.88	0.88	0.88
		Rotten	0.74	0.74	0.74
AM	QuAD	Fresh	0.85	0.92	0.89
		Rotten	0.78	0.63	0.70
	DF-QuAD	Fresh	0.85	0.93	0.89
		Rotten	0.79	0.61	0.69
	REB	Fresh	0.89	0.88	0.88
		Rotten	0.73	0.75	0.74

Table 3: Classification report using SA (682 Fresh and 306 Rotten movies) and AM (620 Fresh and 267 Rotten movies). The difference in the number of movies is due to SA and AM filtering 'neutral' relations using different methods.

Table 2 shows the MAE and RMSE. Since ratings are highly subjective and reviews deemed to be both slightly or highly positive will be classified as Fresh in RT, we focus on the MAE score where individual differences are weighted equally in the average. Using SA, DF-QuAD is the best performing semantics, followed by QuAD and REB. It achieves 15.978 MAE on the scores and 1.527 on the graded scores. Similarly, using AM, DF-QuAD performs best, followed by QuAD and REB. It achieves 16.343 MAE on the scores and 1.573 on the 'graded' scores. The errors for AM are slightly higher than those for SA. One reason for this may be the neural model filtering 'neutral' relations between a sentence from a review and the argument for its associated feature, which, using SA, were deemed to be positive/negative. We also report the Pearson Correlation Coefficient (PCC) which showed similar results across all of the semantics.

The classification report in Table 3 shows less difference between SA and AM, while REB has an advantage w.r.t. F_1.

Table 4 shows examples of the mined sub-features of f_T, an important aspect of ADA. These sub-features can modify a movie's score significantly, e.g. they strengthened *Call Me by Your Name* from 77 to 88 (TS 98) and *The Godfather* from 78 to 90 (TS 95).

From a computational point of view, the strengths can be calculated in polynomial time (since the QBAFs we determine by SA and AM are acyclic). When a new review is added, the pre-trained SA and AM methods simply extend the QBAF. The real computational cost is in training the SA and AM models, which is, however, completed upfront and offline.

Movie	Theme	Keywords
Lady Bird	adolescence	adolescence, coming of age
Get Out	race	black people, of color
La La Land	relationship	rapport, intimacy
Phantom Thread	clothing	clothe, fashion, dress, designer

Table 4: Examples of movies' extracted themes.

5.2 Qualitative Assessment

Focusing on the DF-QuAD algorithm, which showed the smallest error with respect to the TS, we obtain movie scores close to the TS, e.g. *RBG* (86 vs TS 90), exact scores, e.g. *ET The Extraterrestrial* (97), and also results in contrast to the TS, e.g. *The Greatest Showman* (79 vs TS 36). One reason for this is the errors which occur in extracting votes, e.g. from the review: *The director, Michael Gracey, delivers quick doses of excitement in splashy scenes but has little feel for the choreographic action* we only obtain *...doses of excitement in splashy scenes*, resulting in a positive vote. Another is the failure to identify any votes from reviews such as the following (negative) text: *In a broader sense, the mishmash does recall the real Barnum, who once sewed half a fish to half an ape and called it a mermaid*, as it does not mention any (sub-)feature or opinion about the movie itself. With more development of our method and of NLP techniques in general, the frequency and impact of these errors will be reduced.

The difference in the semantics' results in Table 2 can be attributed to the properties that they satisfy. The minor differences between QuAD and DF-QuAD are due to the fact that they only differ in certain situations (see [32]), but this highlights the fact that the property of balance (satisfied by REB and DF-QuAD but not QuAD, see Table 1) may be less crucial in this setting than one might think. The greater accuracy of (DF-)QuAD compared to REB can be explained by the following properties.

The first two properties (forms of strict monotonicity) equate to: adding reasoning against (for) a movie/feature decreases (increases, respectively) its strength. Only REB satisfies these[11].

Property 1. For any $\alpha,\beta,\gamma \in \mathcal{A}$ where $\mathcal{L}^+(\alpha) = \mathcal{L}^+(\beta)$ and $\tau(\alpha)=\tau(\beta)>0$, if $\mathcal{L}^-(\beta)=\mathcal{L}^-(\alpha)\cup\{\gamma\}$ where $\sigma(\gamma)>0$, then $\sigma(\beta)<\sigma(\alpha)$.

Property 2. For any $\alpha,\beta,\gamma \in \mathcal{A}$ where $\mathcal{L}^-(\alpha) = \mathcal{L}^-(\beta)$ and $\tau(\alpha)=\tau(\beta)<1$, if $\mathcal{L}^+(\beta)=\mathcal{L}^+(\alpha)\cup\{\gamma\}$ where $\sigma(\gamma)>0$, then $\sigma(\beta)>\sigma(\alpha)$.

REB satisfying Properties 1 and 2 means that the combined effect of a set of attackers or supporters must never saturate. This means that attackers or supporters will, in general, have less effect on an argument's strength and so its strength is *shifted* towards its base score, when compared with (DF-)QuAD. The semantics may therefore be more suitable for applications with a large number of attackers or supporters, e.g. debates on social media, but here it seems to subdue the attackers and supporters somewhat, e.g. for *Lady Bird*, which has strong supporters, the TS, QuAD, DF-QuAD and REB results are 100, 92, 92 and 85, respectively. We therefore conclude that Properties 1 and 2 (and strict monotonicity in general) characterise behaviour which is not desirable in this setting.

[11] All proofs follow directly by inspection of the semantics' definitions.

The following, novel property, is satisfied by QuAD and DF-QuAD but not by REB. It defines the *attainability* of a semantics, which in our setting equates to: all strength values are attainable for any given base score with a certain set of attackers or supporters.

Property 3. For any $\alpha \in \mathcal{A}$, $\forall v \in [0,1]$, $\exists S, T \in [0,1]^*$ such that if $\sigma(\mathcal{L}^-(\alpha)) = S$ and $\sigma(\mathcal{L}^+(\alpha)) = T$, then $\sigma(\alpha) = v$.

The fact that REB does not satisfy Property 3 highlights a problem in that there is a *blind spot* in its strength results. If we express the algorithm in the form $y = 1 - \frac{1-\tau(\alpha)^2}{1+\tau(\alpha)\cdot e^x}$ we can see that as x approaches ∞ (the balance of reasoning moves in favour of support), $y = 1$, while as x approaches $-\infty$ (the balance of reasoning moves in favour of attack), $y = \tau(\alpha)^2$, i.e. attackers cannot weaken the argument further than the square of its base score. In this context, this implies that as the votes on an argument become stronger, it becomes increasingly resistant to the weakening effect of negative reasoning. This also causes an asymmetry between attackers and supporters, inhibiting the attackers' effect compared to the supporters', which may be suitable in some settings whereas here symmetry is more intuitive. Thus, Property 3 characterises behaviour which is desirable for a semantics in this context.

In Table 3, SA and AM show that REB has a slight advantage over (DF-)QuAD when considering F_1. In terms of Precision, REB is more accurate than (DF-)QuAD for the Fresh class but not for the Rotten class, which is due to REB satisfying Properties 1 and 2 and its shift towards the base score. Given that the Rotten class comprises arguments with strength less than 0.6 and that the midpoint of the base score is 0.5, REB wrongly shifts some movies into the Rotten class while those it classes as Fresh are more likely to be correctly classified. The Recall results back up this finding since REB correctly classifies many more Rotten movies than (DF-)QuAD.

6 DIALOGICAL EXPLANATIONS

The extracted QBAFs also provide the underlying structure for generating dialogical explanations for users. Firstly, we give two properties (reformulations of *Neutrality* [2]), satisfied by all three semantics, stating that attackers/supporters with the minimum strength have no effect on the arguments they attack/support. This allows us to ignore arguments with a base score of 0 and without attackers or supporters, simplifying the explanation aspect of ADA.

Property 4. For any $\alpha,\beta,\gamma \in \mathcal{A}$, if $\tau(\alpha) = \tau(\beta)$, $\mathcal{L}^+(\alpha) = \mathcal{L}^+(\beta)$, $\mathcal{L}^-(\alpha) = \mathcal{L}^-(\beta) \backslash \{\gamma\}$, $\gamma \in \mathcal{L}^-(\beta)$ and $\sigma(\gamma) = 0$ then $\sigma(\beta) = \sigma(\alpha)$.

Property 5. For any $\alpha,\beta,\gamma \in \mathcal{A}$, if $\tau(\alpha) = \tau(\beta)$, $\mathcal{L}^-(\alpha) = \mathcal{L}^-(\beta)$, $\mathcal{L}^+(\alpha) = \mathcal{L}^+(\beta) \backslash \{\gamma\}$, $\gamma \in \mathcal{L}^+(\beta)$ and $\sigma(\gamma) = 0$ then $\sigma(\beta) = \sigma(\alpha)$.

A user may interact with ADA by requesting an explanation of an argument (movie or (sub-)feature).

Definition 6.1. Given any (augmented) review aggregation $\langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ of any $m \in \mathcal{M}$ and corresponding QBAF $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ with strength σ, an *argumentation dialogue* between a user and ADA consists of *explanation requests* $\mathcal{Q}(\alpha)$ for $\alpha \in \mathcal{A}$ from the user, to which ADA responds with *explanations* $\mathcal{X}(\alpha)$.

We define a simple argumentation dialogue as follows.

Definition 6.2. Let $r_a^+, r_a^-, r_b^+, r_b^-$ be functions giving *positive primary*, *negative primary*, *positive secondary* and *negative secondary* (respectively) phrases about an argument, as in Figure 2. For any $S \subseteq \mathcal{A}$, if $S = \varnothing$ let $max(S) = \varnothing$; else, let $max(S) = argmax_{s \in S} \sigma(s)$. Then, a *simple argumentation dialogue* is such that for any $\alpha \in \mathcal{A}$:

$r_a^+(\gamma)$ = {because (the) γ was/were great};
$r_a^-(\gamma)$ = {because (the) γ was/were poor};
$r_b^+(\gamma)$ = {although (the) γ was/were great};
$r_b^-(\gamma)$ = {although (the) γ was/were poor};
$r_a^+(\varnothing) = r_a^-(\varnothing) = r_b^+(\varnothing) = r_b^-(\varnothing) = \{\}$.

Figure 2: Functions $r_a^+, r_a^-, r_b^+, r_b^-$, for any $\gamma \in \mathcal{A}$

if $\alpha = m$ and $\sigma(\alpha) < 0.6$ and $\exists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {Why was α poorly rated?}
$\mathcal{X}(\alpha)$ = {This movie was poorly rated}+
$r_a^-(max(\mathcal{L}^-(m))) + r_b^+(max(\mathcal{L}^+(m)))$; else

if $\alpha = m$ and $\sigma(\alpha) \geq 0.6$ and $\exists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {Why was α highly rated?}
$\mathcal{X}(\alpha)$ = {This movie was highly rated}+
$r_a^+(max(\mathcal{L}^+(m))) + r_b^-(max(\mathcal{L}^-(m)))$; else

if $\alpha \in \mathcal{F}$ and $\mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)$ and $\exists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {Why was/were (the) α considered to be poor?}
$\mathcal{X}(\alpha)$ = {(The) α was/were considered to be poor}+
$r_a^-(max(\mathcal{L}^-(m))) + r_b^+(max(\mathcal{L}^+(m)))$; else

if $\alpha \in \mathcal{F}$ and $\mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)$ and $\exists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {Why was/were (the) α considered to be great?}
$\mathcal{X}(\alpha)$ = {(The) α was/were considered to be great}+
$r_a^+(max(\mathcal{L}^+(m))) + r_b^-(max(\mathcal{L}^-(m)))$; else

if $\mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)$ and $\nexists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {What did critics say about (the) α being poor?}
$\mathcal{X}(\alpha)$ = {[p from $c \in \mathcal{C}$ constituting $\mathcal{V}(c,\alpha) = -$]}; else

if $\mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)$ and $\nexists \beta \in \mathcal{L}^-(\alpha) \cup \mathcal{L}^+(\alpha)$ s.t. $\sigma(\beta) > 0$:
$\mathcal{Q}(\alpha)$ = {What did critics say about (the) α being great?}
$\mathcal{X}(\alpha)$ = {[p from $c \in \mathcal{C}$ constituting $\mathcal{V}(c,\alpha) = +$]}.

Our intuition here is that a simple explanation of the reasoning for each argument's strength may consist of its strongest attacker and its strongest supporter linked by *because*/*although* connectives, depending on whether the argument is Fresh/Rotten for movies or had mostly positive/negative votes for features. If an argument has no attackers or supporters, we use critics' phrases (which constitute the votes towards the argument's base score) to explain its strength. It should be noted that this form of explanation is only possible due to the use of the QBAF mechanism. Other formalisms, or indeed semantics, may lead to different explanations which may be better suited to different contexts. For example the *Propagation Semantics* [7], defined for *Abstract Argumentation Frameworks* (AFs) [13], (restrictions of QBAFs comprising only arguments and attacks) are defined globally, i.e. an indirect (not immediate) attacker can have as much influence as a direct attacker.

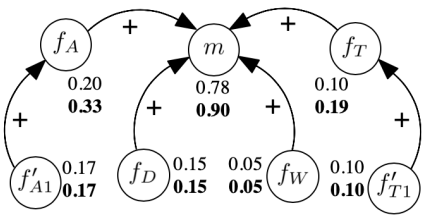

Figure 3: QBAF for the movie *Phantom Thread* with base scores (normal font) and DF-QuAD strengths (bold font).

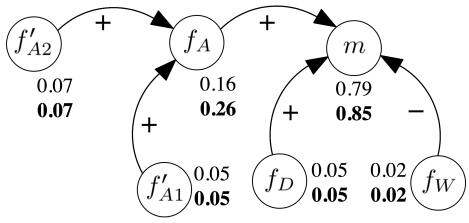

Figure 4: QBAF for the movie *The Post* with base scores (normal font) and DF-QuAD strengths (bold font).

In this case, the main justification may be an argument which is not in the set of direct attackers or supporters.

Consider the QBAF in Figure 3 for the movie *Phantom Thread* (TS=98). All four features support the movie and thus represent positive reasoning for its strength, with *Daniel Day-Lewis* (f'_{A1}) supporting the movie's strongest feature, f_A, and *clothing* (f'_{T1}) supporting the next strongest, f_T. A simple argumentation dialogue between a user and ADA, may then be as follows:

user: *Why was Phantom Thread highly rated?*
ADA: *This movie was highly rated because the acting was great.*
user: *Why was the acting considered to be great?*
ADA: *The acting was considered to be great because Daniel Day-Lewis was great.*
user: *What did critics say about Daniel Day-Lewis being great?*
ADA: *"...Daniel Day-Lewis remains our greatest actor..."*
user: *Why were the themes considered to be great?*
ADA: *The themes were considered to be great because (the theme) clothing was great.*
user: *What did critics say about (the theme) clothing being great?*
ADA: *"...it's set in an evocative ecosphere of haute couture fashion..."*

This dialogue is fairly repetitive given this movie's almost universally positive reviews (as well as our argumentation dialogue's fairly simple nature for illustration) but consider the QBAF for *The Post* (TS=91) in Figure 4. This QBAF has a different structure from that for *Phantom Thread* as no themes were mined for this movie (i.e. f_T has no effect and so is not shown), there is an attack and it has two sub-features supporting f_A. We can see that f_D positively contributed very little and f_W was actually considered to be poor since it attacks m. However, the acting from *Tom Hanks* (f'_{A1}) and, particularly, *Meryl Streep* (f'_{A2}) contributed to the high strength. The argumentation dialogue may then be:

user: *Why was The Post highly rated?*
ADA: *This movie was highly rated because the acting was great, although the writing was poor.*
user: *Why was the acting considered to be great?*
ADA: *Its acting was considered to be great because Meryl Streep was great.*
user: *What did critics say about Meryl Streep being great?*
ADA: *"...Streep's hesitations, rue, and ultimate valor are soul-deep..."*

There are a plethora of ways in which more complicated argumentation dialogues could be defined to give more interesting and varied interactions. For example, the strength scale could be separated to a greater degree and attackers/supporters other than those with the maximum strength could be considered for different levels of sentiment, rather than simply *great*/*poor*, e.g. *the writing was okay but wasn't great*; functions giving phrases could depend upon the type of argument for a more natural sounding phrase, e.g. for an actress: *Meryl Streep put in an excellent performance*; or arguments could be considered in tandem, e.g. *the acting was exceptional thanks to Tom Hanks and, particularly, Meryl Streep.*

7 CONCLUSIONS

We have studied how a popular instance of review aggregation that suffers from a lack of explainability, Rotten Tomatoes' TS, can benefit from NLP and quantitative argumentation integrated within a novel form of (argumentative dialogical) agents that interact with users to provide explanations. The explanations are extracted from a form of argumentation frameworks, QBAFs, mined from review aggregations, in turn mined from the reviews. We experimented with three gradual semantics for QBAFs that give comparable results to the TS. These semantics and properties used in the analysis (one of which is novel and highlights the *blind spot* limitation of a semantics) have been assessed in their suitability to this setting, for both matching the TS and providing explanations to users.

Our method empowers existing techniques (sentiment analysis, argument mining, gradual evaluation for QBAFs) to be deployed, surpassing the individual components it integrates. It is extensible (e.g. we can add more features), and the extensions do not require the models to be retrained (as the training occurs independently of the features). Further, our method is independent of the NLP method used to extract the votes, and as these methods improve over time, our results will do likewise. The advantage of ADA lies in its explanations while giving a comparable measure to the TS.

We foresee numerous directions for future work, from further development to implementation and user studies. ADA could be developed by representing phrases as arguments (see [26]) which, combined with more sophisticated mining techniques, would allow the recognition of similar phrases that can be represented as a single argument or the detection of attacks/supports between phrases (see [9]). We envisage developing ADA to utilise both NLP methods (SA and AM) together in a complimentary manner, rather than comparing the two methods separately. ADA could be tested on larger datasets, e.g. the full RT reviews (as opposed to the snippets), while user studies on the best way to provide the explanations, e.g. graphical, visual or linguistic, for different contexts and users, as in [17], would also be fruitful. Further work on generating dialogical explanations, e.g. using restrictions of QBAFs or different semantics, could lead to a more evolved dialogue between ADA and users.

REFERENCES

[1] Leila Amgoud and Jonathan Ben-Naim. 2017. Evaluation of Arguments in Weighted Bipolar Graphs. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 14th European Conference, ECSQARU*. 25–35.

[2] Leila Amgoud and Jonathan Ben-Naim. 2018. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning* 99 (2018), 39–55.

[3] Katie Atkinson, Trevor J. M. Bench-Capon, and Peter McBurney. 2005. A Dialogue Game Protocol for Multi-Agent Argument over Proposals for Action. *Autonomous Agents and Multi-Agent Systems* 11, 2 (2005), 153–171.

[4] Pietro Baroni, Antonio Rago, and Francesca Toni. 2018. How Many Properties Do We Need for Gradual Argumentation?. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 1736–1743.

[5] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6, 1 (2015), 24–49.

[6] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better Document-level Sentiment Analysis from RST Discourse Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 2212–2218.

[7] Elise Bonzon, Jérôme Delobelle, Sébastien Konieczny, and Nicolas Maudet. 2016. Argumentation Ranking Semantics Based on Propagation. In *Computational Models of Argument - Proceedings of COMMA 2016*. 139–150.

[8] Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. In *Computational Models of Argument COMMA*. 21–32.

[9] Lucas Carstens and Francesca Toni. 2017. Using Argumentation to Improve Classification in Natural Language Problems. *ACM Transactions on Internet Technology* 17, 3 (2017), 30:1–30:23.

[10] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications* 72 (2017), 221–230.

[11] Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1374–1379.

[12] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*. 484–495.

[13] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–358.

[14] Xiuyi Fan, Francesca Toni, Andrei Mocanu, and Matthew Williams. 2014. Dialogical two-agent decision making with assumption-based argumentation. In *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS*. 533–540.

[15] Stefano Ferilli, Andrea Pazienza, Sergio Angelastro, and Alessandro Suglia. 2017. A Similarity-Based Abstract Argumentation Approach to Extractive Text Summarization. In *AI*IA 2017 Advances in Artificial Intelligence - XVIth International Conference of the Italian Association for Artificial Intelligence*. 87–100.

[16] Alejandro Javier García, Carlos Iván Chesñevar, Nicolás D. Rotstein, and Guillermo Ricardo Simari. 2013. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications* 40, 8 (2013), 3233–3247.

[17] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[18] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

[19] Emmanuel Hadoux, Anthony Hunter, and Jean-Baptiste Corrégé. 2018. Strategic Dialogical Argumentation Using Multi-criteria Decision Making with Application to Epistemic and Emotional Aspects of Arguments. In *Foundations of Information and Knowledge Systems - 10th International Symposium, FoIKS*. 207–224.

[20] Gauri Jain, Manisha Sharma, and Basant Agarwal. 2018. Spam Detection on Social Media Using Semantic Convolutional Neural Network. *IJKDB* 8, 1 (2018), 12–26.

[21] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology* 16, 2 (2016), 10.

[22] Peter McBurney, Rogier M. van Eijk, Simon Parsons, and Leila Amgoud. 2003. A Dialogue Game Protocol for Agent Purchase Negotiations. *Autonomous Agents and Multi-Agent Systems* 7, 3 (2003), 235–273.

[23] Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. 2014. Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews. (2014). http://arxiv.org/abs/1412.5335

[24] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the Association for Computational Linguistics (ACL)*. 271–278.

[25] Viraj Parkhe and Bhaskar Biswas. 2016. Sentiment analysis of movie reviews: finding most important movie aspects using driving factors. *Soft Computing* 20, 9 (2016), 3373–3379.

[26] Theodore Patkos, Antonis Bikakis, and Giorgos Flouris. 2016. A Multi-Aspect Evaluation Framework for Comments on the Social Web. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR*. 593–596.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1532–1543.

[28] Henry Prakken. 2005. Coherence and Flexibility in Dialogue Games for Argumentation. *Journal of Logic and Computation* 15, 6 (2005), 1009–1040.

[29] Henry Prakken. 2006. Formal systems for persuasion dialogue. *Knowledge Engineering Review* 21, 2 (2006), 163–188.

[30] Antonio Rago, Oana Cocarascu, and Francesca Toni. 2018. Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*. 1949–1955.

[31] Antonio Rago and Francesca Toni. 2017. Quantitative Argumentation Debates with Votes for Opinion Polling. In *PRIMA*. 369–385.

[32] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *KR*. 63–73.

[33] Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*. 438–449.

[34] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification.. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1422–1432.

[35] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. 2016. Classification of Sentiment Reviews Using N-gram Machine Learning Approach. *Expert Systems with Applications* 57, C (Sept. 2016), 117–126.

[36] Azene Zenebe and Anthony F. Norcio. 2009. Representation, Similarity Measures and Aggregation Methods Using Fuzzy Sets for Content-based Recommender Systems. *Fuzzy Sets anf Systems* 160, 1 (Jan. 2009), 76–94.

[37] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase-level Sentiment Analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. 83–92.